Image Steganography
based on Iterative Adversarial perturbations
onto A Synchronized-directions Sub-image

Xinghong Qin, Shunquan Tan, Weixuan Tang,
Bin Li and Jiwu Huang

Shenzhen University
Introduction

- Steganography and steganalysis are a pair of antagonistic players.
  - Steganography:
    - Steganography is trying to escape being detected by steganalysis.
  - Steganalysis:
    - The warden discriminates whether a cover or a stego object is sent.

- Scenario
  - The sender slightly modifies the cover $C$ to conceal the secret message $M$ to produce the stego $S$.
  - Send $S$ to the receiver through the channel with passive the warden.
  - The receiver extracts $M$ from the received $S$.
  - If the warden classifies the sent object is a stego, he maybe block-up the transmission or damage the sent object.
Introduction

• Steganography has to face challenges of both feature-based steganalysis and CNN steganalysis.
Introduction

• **Motivation.**
  
  – Incorporate **SMD strategy and adversarial examples** to further enhance steganographic security to counter both feature-based steganalysis and CNN steganalysis.
  
  • Synchronizing modification directions (SMD) strategy can improve steganographic security.

\[
D(X, Y) = \sum_{(i,j), (k,l) \in \mathcal{C}} S_C(X_{ij} - Y_{ij}, X_{kl} - Y_{kl})
\]  

(1)

• Many machine learning classifiers are vulnerable to adversarial examples.

\[
X_{adv} = X + \epsilon \cdot \text{sign}(\nabla_X \mathcal{L}(\Phi(X), y_t))
\]  

(2)

**ITE-SYN: Xinghong Qin, Shunquan Tan, Weixuan Tang, Bin Li and Jiwu Huang. IEEE ICASSP 2021**
Our Method

- Base framework
  - ITE-SYN: \textit{ITE}ratively apply adversarial perturbations onto one \textit{SYN}chronized modification directions sub-image.

\begin{itemize}
  \item ITE-SYN
  \begin{itemize}
  \item Embed secret message with synchronizing modification directions
  \item Iteratively apply adversarial perturbations onto one sub-image
  \end{itemize}
\end{itemize}
Our Method

- Embed secret message with synchronizing modification directions
  - Implement clustering modification directions (CMD) strategy.
    - The initial costs $\xi$ are only computed once.
    - Adjust costs as
      \[
      \begin{align*}
      \rho_{+}^{(i,j)} &= \frac{\xi_{+}^{(i,j)}}{\beta} & & \text{if} & \sum_{\Delta e^{(r,s)} \in N^{(i,j)}} \Delta e^{(r,s)} > 0, \\
      \rho_{-}^{(i,j)} &= \frac{\xi_{-}^{(i,j)}}{\beta} & & \text{if} & \sum_{\Delta e^{(r,s)} \in N^{(i,j)}} \Delta e^{(r,s)} < 0, \\
      \rho_{\pm}^{(i,j)} &= \xi_{\pm}^{(i,j)} & & \text{otherwise},
      \end{align*}
      \]
      where $N^{(i,j)} = \{\Delta e^{(r,s)} | r \in \{i - 1, i + 1\}, s \in \{j - 1, j + 1\}\}$
      \[\Delta C = S - C\]
    - Select $\beta = 10$ for images with size-scale $256 \times 256$.
Our Method

• Iteratively apply adversarial perturbations.
  – We re-embed image to produce adversarial perturbations.
    \[
    \Delta C' = Z - C = (Z - S) + (S - C) = n + \Delta C,
    \]
  – Adversarial costs are computed from embedding costs \( \rho \) adjusted by SMD.
    \[
    \rho^{(i,j)}_{adv+} = \begin{cases} 
    \rho^{(i,j)}_+ (1 + \gamma) & \text{if } \nabla \mathcal{L}^{(i,j)}(S, y_c) > 0, \\
    \rho^{(i,j)}_+ / (1 + \gamma) & \text{if } \nabla \mathcal{L}^{(i,j)}(S, y_c) < 0, \\
    \rho^{(i,j)}_+ & \text{otherwise}
    \end{cases}
    \]
    \[
    \rho^{(i,j)}_{adv-} = \begin{cases} 
    \rho^{(i,j)}_- (1 + \gamma) & \text{if } \nabla \mathcal{L}^{(i,j)}(S, y_c) < 0, \\
    \rho^{(i,j)}_- / (1 + \gamma) & \text{if } \nabla \mathcal{L}^{(i,j)}(S, y_c) > 0, \\
    \rho^{(i,j)}_- & \text{otherwise}
    \end{cases}
    \]

  – Parameters
    \[
    \Delta \gamma = 0.1, \\
    \gamma_{max} = 10.
    \]
Our Method

- Iteratively apply adversarial perturbations.
  - Adversarial perturbations are **only** applied onto one sub-image.
  - If re-embedding one sub-image is failed to deceive the target CNN classifier, the next sub-image will be selected to be re-embedded until all sub-images are tried re-embedding.
Experiments

• Setup
  – Image database: BOSS256
    • Union of BOSSBase v1.01 and BOWS2. Totally 20000 images.
    • Resize each image from size-scale 512X512 to 256X256 by Matlab.
    • For CNN, 1000 images and 5000 images randomly selected from BOSSBase for validation and testing, other 14000 images are for training.
  – Cost functions
    • Heuristic method: HILL.
    • Model-based method: MiPOD.
  – Steganalysis
    • CNN classifiers
      – The target: XuNet, YeNet.
      – The non-target: SRNet.
    • Ensemble classifiers: SRM, maxSRMd2, PDASS.

• Comparison schemes
  • ADV-EMB
  • MinMax + ADV-EMB.

• Payload rates
  • 0.2 bpp and 0.4 bpp

• Performance
  \[ P_E = \frac{P_{FA} + P_{MD}}{2} \]  

• Stegos are created by the simulator unless specified.
Experiments

- Deceiving original classifiers
  - Notations
    - BAS: baseline.
    - ADV: ADV-EMB.
    - ITE: ITE-SYN.
    - M1-M9: versions of MinMax+ADV-EMB.
  - Target CNN classifier
    - XuNet: (a)-(b)
    - YeNet: (c)-(d)

- Conclusion
  - ITE-SYN can effectively deceive the target CNN classifiers.
  - ITE-SYN improve steganographic performances to counter other original classifiers.
Experiments

- Countering adversarial training classifiers
  - ITE-SYN outperforms ADV-EMB.
  - For comparison with MinMax+ADV-EMB,
    - ITE-SYN performs superior for non-target CNN classifiers and feature-based classifiers.
    - ITE-SYN performs superior when countering YeNet classifiers.
    - MinMax+ADV-EMB outperforms ITE-SYN after the fourth round when countering XuNet classifier.

- Discussion
  - Computational complexity of ITE-SYN is lower than of MinMax+ADV-EMB.
    - ITE-SYN creates only one stego image for each cover image.
  - It is predicted that steganographic performances of MinMax+ITE-SYN should be further improved.
Appendix: Issues

• Performances of MinMax+ITE-SYN
  – Notations
    • BAS: baseline.
    • M0-M9: rounds of MinMax.
  – Conclusion
    • MinMax+ITE-SYN outperforms MinMax+ADV-EMB.

Performances of countering adversarial training classifiers.

ITE-SYN: Xinghong Qin, Shunquan Tan, Weixuan Tang, Bin Li and Jiwu Huang. IEEE ICASSP 2021
Experiments

- **Computational time (STCs)**
  - Success rates are over 90%.
    - ITE-SYN can effectively deceive the target CNN classifiers.
  - Maximal iteration
    - ADV-EMB: 10.
    - ITE-SYN: 400.
  - Average computational times of ITE-SYN are less than of ADV-EMB, except for ITE-SYN for XuNet with payload rate 0.2 bpp.
    - Success rate of ITE-SYN is less about 5%.
- **Cumulative success rate**
  \[ P_\gamma(x_0) = \int_{-\infty}^{x_0} f(t) dt \]
  - When \( \gamma_{\text{max}} = 1 \)
    - cumulative success rates are over 80%,
    - the maximal iteration of ITE-SYN: 40,
    - average time of creating adversarial stego image by ITE-SYN for XuNet as the target CNN classifiers with payload rate 0.2 bpp is 7.38 seconds.

### Table: Average success rate and computational time

<table>
<thead>
<tr>
<th>Target</th>
<th>Scheme</th>
<th>0.2 bpp</th>
<th>0.4 bpp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Success rate</td>
<td>Time</td>
</tr>
<tr>
<td>XuNet</td>
<td>ADV-EMB</td>
<td>95.76</td>
<td>10.40</td>
</tr>
<tr>
<td></td>
<td>ITE-SYN</td>
<td>90.79</td>
<td>24.19</td>
</tr>
<tr>
<td>YeNet</td>
<td>ADV-EMB</td>
<td>99.82</td>
<td>6.68</td>
</tr>
<tr>
<td></td>
<td>ITE-SYN</td>
<td>98.95</td>
<td>6.66</td>
</tr>
</tbody>
</table>

- **Conclusion**
  - Computational complexity of ITE-SYN is lower.

ITE-SYN: Xinghong Qin, Shunquan Tan, Weixuan Tang, Bin Li and Jiwu Huang. IEEE ICASSP 2021
Conclusion

• ITE-SYN further enhances steganographic security countering both feature-based steganalysis and CNN steganalysis.
  – ITE-SYN can effectively deceive the target CNN classifiers, and can effectively resist on detection of other original classifiers, including both feature-base classifiers and CNN classifiers.
  – ITE-SYN has significant undetectability to counter adversarial training classifiers, including both feature-based classifiers and CNN classifiers.
  – Gradually increased adversarial perturbations are only applied onto one clustering modification directions sub-image.
    • It spends low computational expense.
    • It guarantees that adversarial perturbations applied are minimal.
    • It is unnecessary to search the optimal adversarial intensity.

• Future works
  – Extend the method to JPEG images.
    • Investigate incorporation of adversarial perturbations and effective cost strategy.
  – Investigate inner mechanisms of both SMD strategy and adversarial perturbations to design more powerful steganographic algorithm.
Thanks!